

"Express Mail" mailing label number:

EL151293368US

## A SWITCHED NETWORK FOR LOW LATENCY COMMUNICATION

Hans Eberle  
Neil C. Wilhelm

### **BACKGROUND OF THE INVENTION**

#### 5 **Field of the Invention**

The present invention relates to networks, and, more particularly, to a switched network.

#### **Description of the Related Art**

Computer networks are an increasingly important part of both private and  
10 business environments. Computing devices such as workstations, personal  
computers, server computers, storage devices, firewalls and other computing devices  
function as nodes of a network with at least one network element connecting the  
computing devices. The various nodes transmit and/or receive various kinds of  
information over the network. The network may be bus based, ring based, a switched  
15 network or a combination. Computing devices and users are demanding higher  
communication speeds across networks as more and more information flows across  
the various networks. The introduction of new technologies will likely load down  
networks even more.

There are numerous network architectures used to interconnect the various  
20 nodes. One common and familiar architecture is a local area network (LAN), which is  
generally a network within a single building or company campus. The rules by which  
nodes transmit and receive packet data are defined in various protocols. One common  
protocol utilized by LANs is defined in IEEE 802.3, also referred to as the Ethernet.  
Other protocols commonly utilized are ring-based protocols such as IEEE 802.5,  
25 referred to as a "token ring" protocol, which requires a special bit pattern, or "token"  
to circulate when nodes are idle, and which nodes remove before transmitting data  
packets.

5 A network protocol provides rules to route a packet of information from a source to a destination in a packet switching network. A packet is generally a portion of a message transmitted over a network that typically includes routing or destination information in addition to data information. Packets may vary in size from only a few bytes to many thousands of bytes.

10 The network protocol acts to control congestion when a resource conflict arises. Resource conflicts arise when network resources are simultaneously requested. The Ethernet (IEEE 802.3), for example, uses a bus-based broadcasting mechanism that allows nodes to transmit at any time. That can result in collisions on the bus. If, in Ethernet based networks, two or more packets collide, the nodes wait a random amount of time before re-transmitting. The sending node typically buffers packets until they are acknowledged because the packets might have to be retransmitted. Receiving nodes may also buffer packets.

15 The type of networks typically used for LANs however, cannot adequately support systems requiring low forwarding latencies and high communication bandwidth, such as distributed processing systems, in which storage resources as well as processing tasks may be shared.

20 In switched networks, similar considerations apply. In a switched network delays occur in the switches when congestion causes packets to be temporarily stored in buffer memories. Congestion arises when a path, internal or external to the switch, is requested to forward more packets than its capacity allows. Usually, it cannot be predicted how long congestion lasts. Thus, forwarding delays are variable and unpredictable. That complicates network design, in particular, it complicates the bookkeeping of outstanding packets and the scheduling of the network switches.

25 Bookkeeping is complex since the number of outstanding packets can vary and since it can be difficult to decide whether a packet was lost or just delayed for a long time. Scheduling the switches is complicated since the routes of the packets cannot be known before the packets actually arrive making it necessary to calculate the routes "on the fly".

30 Another factor to be considered in trying to achieve an efficient network is that data transfers across most networks typically have wide variation in bandwidth

09540730-033100

and latency requirements. Latency and bandwidth define the speed and capacity of a network. Latency is generally the amount of time it takes for a packet to travel from its source to its destination. Bandwidth is the amount of traffic a network can carry in a fixed time, typically expressed in bytes per second. There can be conflicts between a desire for high bandwidth and low latency. For example, in a high speed data network that generally carries large sized data packets (e.g., 2K bytes), a small packet (e.g., 64 bytes) having low-latency requirements, can wait a long time for a large packet currently being transferred to complete. High-bandwidth network traffic with larger-sized packets can conflict with low-latency traffic with smaller-sized packets. Larger-sized packets increase the latency of smaller-sized packets, and smaller-sized packets can interfere with scheduling for larger-sized packets. The smaller-sized packets can prevent larger packets from fully utilizing available bandwidth.

It would be desirable to reduce complexity of network design by avoiding forwarding delays that are variable and unpredictable, avoid complicated bookkeeping related to outstanding packets and scheduling. It would also be desirable to accomplish reduced complexity and still provide higher throughput and it would be desirable to provide a network that addressed the conflict between low latency and high bandwidth traffic. It would also be desirable to reduce complexity of network design by providing a simple switch implementation in which overhead is reduced by simplifying assumptions about availability of switch resources. It would also be desirable to provide such a switch that accommodated low latency traffic effectively.

### **SUMMARY OF THE INVENTION**

Accordingly, in one embodiment, the invention provides a method for communicating packets between one of a plurality of sending nodes and one of a plurality of receiving nodes of a switched network. The switched network includes a buffer-less switch coupling the sending nodes and the receiving nodes. The method includes transmitting packets from respective sending nodes to respective input ports of the buffer-less switch and forwarding all packets that are successfully delivered to the receiving nodes through output ports of the buffer-less switch, through the buffer-less switch with a fixed forwarding rate. The switched network resolves conflicts in requests for a transmission path on the switch from multiple packets by allocating the

transmission path to a first requester for the transmission path. If multiple requests for a switch resource collide by requesting the switch resource at the same time, one of the requests is selected as a winner and the one or more remaining packets associated with the losing request(s) are dropped. The winning packet may be selected  
5 on a random basis or a round robin basis or based on some other criteria. The requests for transmission paths may be contained within the packets sent into the network and extracted on entry of the packet into the network. A sending node can check for successful transmission of a packet at a predetermined time because a receiving node sends an acknowledge back after a fixed delay. A time out can also be  
10 used to determine if a packet was transmitted successfully by checking after the predetermined time for the acknowledge.

In another embodiment the invention provides a method for utilizing a switch of a switched network that includes forwarding packets on a first come first served basis from respective input ports at a fixed rate through the switch to respective output  
15 ports. If a first and second packet simultaneously request a switch resource, one of the first and second packets is designated a winner and one a loser, the winner obtaining the switch resource, and the loser is dropped.

In still another embodiment of the invention a computing system is provided that includes a plurality of sending and receiving nodes, each of the sending and  
20 receiving nodes including a processor. The computing system also includes a low latency switched network including a buffer-less switch coupling the plurality of sending and receiving nodes. The buffer-less switch has a fixed forwarding delay for all packets sent from one of the sending nodes and successfully received by one of the receiving nodes because no buffering takes place within the switch in case of  
25 contention for resources. The switch includes switch control logic, coupled to input registers in the switch. The switch control logic is responsive to packet information stored in the registers, to allocate output ports on the switch according the packet information on a first come first served basis. Respective packet information provided to the switch control logic constitutes respective requests for output ports,  
30 and if a first and second request for an output port path collide by requesting the

output port at the same time, the switch control logic responds by selecting one of the requests as a winner and dropping the other packet.

# **BRIEF DESCRIPTION OF THE DRAWINGS**

The present invention may be better understood, and its numerous objects, features, and advantages made apparent to those skilled in the art by referencing the accompanying drawings.

**FIG. 1** is a block diagram showing a data network with two transmission channels.

**FIG. 2** is a block diagram of the data structures used by a network interface.

**FIG. 3** is a block diagram of a switch suitable for an embodiment of the present invention.

**FIG. 4** is a block diagram of a representative network including two bufferless switches and a switch scheduler and a plurality of network nodes according to an embodiment of the present invention.

**FIG. 5** is a block diagram illustrating a bufferless switch in accordance with an embodiment of the present invention.

**FIG. 6** is a block diagram illustrating aspects of a network node according to an embodiment of the present invention.

**FIG. 7** is a block diagram illustrating aspects of a network node according to an embodiment of the present invention.

**FIG. 8** is a block diagram of a simple 2X2 switch that may be used to implement the low latency switch.

**FIG. 9A** illustrates that a first in time packet wins, in accordance with one embodiment of the low latency switch.

**FIG. 9B** illustrates an embodiment of the low latency switch where one packet is chosen as the winner based on a simple algorithm.

**FIG. 10** is a block diagram of one embodiment of the lossy network.

**FIG. 11A and 11B** are diagrams illustrating advantages of a pipelined network.

**FIG. 12** is a diagram illustrating the various stages for several operations taking place on a pipelined network.

**FIG. 13** is a diagram of a pipelined network in which collision avoidance and detection techniques can be utilized.

**FIG. 14** is a diagram illustrating collision avoidance techniques in a pipelined network.

**FIG. 15** is a diagram illustrating operation of collision detection techniques in a pipelined network.

**FIG. 16** illustrates a multi-stage switch configuration.

The use of the same reference symbols in different drawings indicates similar or identical items.

## **DESCRIPTION OF THE PREFERRED EMBODIMENT(S)**

Referring to Figure 1, a block diagram describes a data network system according to an embodiment of the present invention. Data network system **100** is a network separated into at least two channels **130** and **140**. The two channels **130** and **140** are separate physically and logically. Data network system **100** includes a plurality of nodes **150**, **160**, **180**, and **190** coupled to channels **130** and **140**. Although four nodes are shown, one of skill in the art appreciates that the number of nodes is not limited to four, and may be altered according to system design requirements. Each of nodes **150**, **160**, **180** and **190** optionally are computing devices, such as workstations, personal computers, and server-type computers or other devices that may be coupled to a network such as storage devices and input/output devices. The nodes may be coupled into a distributed computing system through channels **130** and **140**.

15

30

Desired bandwidth and latency characteristics of packets are only examples of characteristics which can be used to select a channel for transmission. Packets may be selected for transmission across one of the channels according to various criteria such as size of a data information packet, type of operation associated with the data information packet, a latency budget for the data information packet, security needs of the data information packet, reliability needs of the data information packet, as well as scheduling strategies of the various channels, e.g., highly scheduled versus limited scheduling, buffering requirements, and error parameters.

Channels can be optimized to carry traffic based on the various criteria in addition to bandwidth and latency. That is, channels can be designed to transport traffic having one or more of the above described criteria. Thus, if other criteria, such as reliability or security are being used, the channels may be optimized differently from the high bandwidth channel and the low latency channel to accommodate such traffic. For example, for traffic having higher reliability needs, a channel can be designed to include a forward error correction scheme that can detect and correct a significant number of expected errors. Thus, an important transfer, e.g., reconfiguration information, may be assigned to the most reliable channel. For simpler reliability needs, a channel can use parity, a checksum, or a cyclic redundancy check (CRC) scheme to detect errors. In addition, security concerns may be addressed by providing a channel that is more physically secure, providing, e.g., detection capability if security of the channel has been compromised. In addition, more complex encryption algorithms may be utilized on a channel designed to accommodate traffic with higher security needs. The channels can of course be designed to carry traffic having one or more of the criteria described herein. For example, a high bandwidth channel may also be designed to provide higher security.

Each of channels **130** and **140** schedule transmissions of data packets through data network system **100** according to requirements of the respective identified features of groups of data packets. Channel **130**, which is designed to transmit low latency packets, uses limited scheduling because an efficient channel transmitting low latency packets requires quick scheduling decisions. Additionally, low latency packets are typically smaller-sized packets that do not cause long lasting blockages. The transmission error rate, therefore, may be of less concern for low-latency channel



130 because an error affects a relatively short data transfer. Therefore, retransmission of a packet that had a transmission error has an acceptable overhead.

On channel 130, the scheduling may be accomplished by allocating a transmission path across the network as the packets arrive in the data network.

5 Assuming a switched data network, the packet or packets may be transmitted to a switch, whereupon switch control logic allocates a transmission path through the switch. The transmission path information, i.e., a desired destination, is typically contained in the packet, commonly in a header or first few bytes of the packet. At the input to the switch the header information is provided to appropriate switch control  
10 logic, which allocates a transmission path to the data packet associated with the transmission path.

As described, channel 140 carries high bandwidth data packets. To maximize the bandwidth, channel 140 operates with more scheduling. In contrast to low latency channel 130, channel 140 is carefully scheduled to maintain a constant flow of data  
15 packets. Channel 140 is designed for transmitting larger-sized packets that can cause longer lasting blockages and that can tolerate increased latency. Longer packets generally have lower overhead than shorter packets on a per byte basis. Therefore, channel 140 has a higher effective throughput of information. Additionally, channel 140 preferably may have a lower error rate than would be acceptable on channel 130.  
20 That is because an error on channel 140 typically affects a relatively large data transfer causing considerable overhead in case retransmission of a packet is required.

High-bandwidth channel 140, which may be scheduled more carefully than low-latency channel 130, can be scheduled prior to transmitting data packets to the data network. Assume the selection criteria determining over which channel to  
25 transmit data is based on data packet size. For those packets that are determined to meet the size criteria, the packets are transmitted with a high degree of scheduling to ensure high utilization of channel 140. The channel transmitting the larger sized data packets may be a highly scheduled channel, a synchronous channel, a pipelined channel, or a channel having those or any combination of those qualities suited for  
30 transmitting larger sized data packets as discussed herein.

09540730-033100

The dual channel architecture described herein is particularly well suited to meet the communication needs of a cluster. A cluster is a group of servers or workstations that work collectively as one logical system. One advantage of clustering is high availability and high performance. Clusters capitalize on economies of scale and are inexpensive alternatives to other fault tolerant hardware-based approaches as well as to other parallel systems such as symmetric multi-processors, massively parallel processors and non-uniform memory architecture machines. The dual channel architecture described herein can guarantee low latency, even under heavy load. Low latency facilitates tight coupling between the nodes of a cluster.

One way to increase efficiency of the system illustrated in Fig. 1 with relation to use of high bandwidth channel 140 is illustrated in Fig. 2. According to one embodiment, channel 140 allocates resources prior to allowing data packets to leave their respective nodes. Sending node 150 and receiving node 160 each set up transfer descriptors 170. As shown in Figure 2, transfer descriptors 170 point to linked lists of memory segment descriptors 210, which include an address descriptor 220 and a length descriptor 230. The address and length descriptors provide a starting address and the length of the memory segment 250 located in memory 240. Each sending node 150 and receiving node 160 sets up transfer descriptors 170 prior to transferring data packets into the data network system. Thus, after a transfer begins, which may involve multiple data packets, data to be sent to receiving node 160 can efficiently be gathered from memory 240 within the sending node 150, and data that is received from the network can efficiently be delivered to memory 240 within the receiving node 160 according to transfer descriptors 170.

As described above, in one embodiment, packet size provides one of the criteria used to select whether traffic should be transmitted over low latency channel 130 or high bandwidth channel 140. Large packets are transferred over one transmission channel, a high bandwidth channel, and small packets are transferred over another transmission channel, a low latency channel. The sending node determines whether a particular packet should be transferred over low latency channel 130 or high bandwidth channel 140. The exact criteria for whether a packet is considered large or small depends on system design requirements. For example, a particular system may require that the transfer be of at least a predetermined threshold

5

10

15

25

5

10

15

20

25

30

According to an embodiment of the present invention, one of the switches is a non-blocking buffer-less switch. Alternatively, each of channels 130 and 140 uses switches that may or may not be buffer-less and may or may not be blocking-type switches. In an exemplary embodiment, the switches are configured according to the channel requirements. For example, a channel optimized to transmit highly scheduled high bandwidth packets includes a non-blocking buffer-less switch, as more fully described below. A channel optimized to transmit low latency data optionally may include a switch that allows blocking of packets.

One type of switch appropriate for an embodiment is shown in Figure 3.

Referring to Figure 3, a block diagram shows a crossbar switch 300. Each of input ports 310 is coupled to each of output ports 320. Assuming each input port 310 and each output port 320 have the same bandwidth "b," resource conflicts can arise. According to an embodiment, if no buffer memory is present in the switch 300 to temporarily store data packets, and multiple data packets are simultaneously forwarded to one of output ports 320, switch 300 drops data packets.

One method of preventing conflicts requires an input buffer memory or output buffer memory to temporarily store packets. An input buffer holds a data packet in an input buffer coupled to the switch 300 and prevents the data packet from entering the switch 300 until a desired one of the output ports 320 is available. Similarly, output buffering avoids conflicts by providing an output buffer memory with enough input bandwidth to allow packets to be received simultaneously from all input ports 310. One or more channels using a switch with input or output buffers is within the scope of the present invention.

Referring now to Figure 4, a block diagram illustrates an exemplary switched data network embodiment employing two buffer-less switches, each switch transmitting packets for a different type of channel. In the embodiment, the switches are coupled to switch scheduler 440. In the embodiment, a channel for transmitting high bandwidth, larger-sized packets is represented by high bandwidth or bulk channel switch 450, which may be a flow-through switch. A channel for transmitting low latency, smaller-sized packets is represented by low-latency or quick channel switch 460.

5

10

15

25

5 This embodiment is also suitable for configurations in which the bulk channel switch 450 has an optical interconnect or an optical switch or both, which may make transfer of control information difficult. Using a separate channel for routing control information allows the bulk channel to benefit from the higher speeds of an optical configuration. In addition, if an optical interconnect and switch are utilized for both  
10 the bulk and quick channel, wavelength may be used to distinguish a low latency channel from a high bandwidth channel in addition to distinguishing the output ports.

In one embodiment, as described further herein, quick channel switch 460 has minimum scheduling overhead and no buffering, resulting in dropping of packets when collisions occur. The lossy nature of the quick channel in such an embodiment could lead to unwanted loss of request and grant packets resulting in loss of bulk channel bandwidth. However, request and grant packets are treated in a manner that avoids such dropping. More particularly, request packets are forwarded directly from the input ports 422 of quick channel switch 460 to the switch scheduler 440 without passing through the switching fabric of quick channel switch 460 (i.e., without passing through the output ports connected to the other nodes). The scheduler 440 is

capable of receiving request packets from each of the nodes 420 simultaneously. That configuration avoids collisions within the switching fabric and the potential of dropping request packets.

Conversely, the switch scheduler 440 transmits grant packets generated in the arbitration logic within the switch scheduler 440 to output ports 432 of the quick channel switch 460. The grant packets may collide with other packets that are simultaneously forwarded to the output ports of the quick channel. Due at least in part to the important nature of the grant packets for scheduling the bulk channel switch 450, the grant packets are prioritized in the event of a collision. Thus, if a collision with a grant packet occurs in quick channel switch 460, the grant packets are given higher priority and are forwarded and other packets are dropped. The quick channel switch 460 sends the grant packets simultaneously to all nodes at a predetermined time within a bulk frame time period. That predetermined time is known by all nodes in the network. Thus, the nodes can avoid collisions with the grant packets by avoiding transmittal of any packets during the time periods predetermined to be assigned to grant packets, to better optimize use of quick channel 460.

If it is desired to minimize wire and pin counts, quick channel switch 460 may be implemented as a serial switch, in which either the ports and/or the internal data paths through the switch are serial. Bulk channel switch 450 may also be realized as a switch in which ports as well as internal data paths are serial. In other implementations one or both of the ports and internal data paths of bulk channel switch 450 may be parallel. Note that in one embodiment bulk channel switch 450 does not need to resample data and can be realized as a switch with all combinational logic (e.g. multiplexers). That is, it has no clocked logic in the form of buffers or registers.

Many different arbitration schemes may be utilized to schedule the bulk channel. In one embodiment, the arbitration scheme allocates output ports as a function of the number of requests being made by an input port. Those input ports making the fewest requests are scheduled first. In another embodiment, the arbitration scheme may allocate output ports based on the number of requests being



made for a particular output port. Those output ports with the fewest requests are allocated first. A round robin scheme can also be used by the arbiter to avoid starvation in conjunction with those embodiments. Further details on an arbiter which may be used in some or all of the embodiments described herein, are described in the patent application entitled "Least Choice First Arbiter", naming Nils Gura and Hans Eberle as inventors, application number \_\_\_\_\_ (Attorney Docket Number 1004-4282), filed the same day as the present application and which is incorporated herein by reference. Of course, one of ordinary skill would understand that many other arbiters are known in the art and may be utilized in the various embodiments described herein.

In an embodiment, a precalculated schedule is established before the bulk scheduler/arbiter does its work. It is precalculated either by one of the nodes in the form of a centralized scheduler or by all the nodes in the form of a distributed scheduler.

The precalculated schedule may be used to implement quality of service (QoS), e.g., transmission of audio or video streams. The source of the stream asks the scheduler to periodically reserve a switch slot. For example, if the link bandwidth is 2.5 Gbits/s and the stream requires a bandwidth of 2.5 Mbytes/s, the source of the stream asks the scheduler to reserve 1 slot every 1000 bulk frames.

The precalculated schedule may be communicated to the bulk scheduler with the help of the request packets. For every slot on bulk channel switch the scheduler receives one request packet from every node. That request packet contains an additional vector of prescheduled targets. The bulk scheduler uses that information in that the scheduler does not schedule the output ports that are already reserved by the precalculated schedule. While the precalculated schedule is required to be conflict-free, the bulk scheduler does check whether this is the case to ensure that collisions are avoided due to an erroneous precalculated schedule.

The precalculated schedule allows for multicast. That is one reason why the request packet contains a vector. The vector specifies to which target or targets the initiator will send a bulk packet.

10

20

30

According to another embodiment, each network node **520, 530, 540, and 550** includes multiple storage queues. Thus, each network node includes a queue for sending packets and a queue for receiving packets, or, alternatively, one or more send queues and receive queues. Thus, each input port couples to a queue and each output port couples to a queue. Each queue disposed within each network node may include a portion of the queue within a network interface. Advantageously, having multiple send queues provides more choice when establishing connectivity between input ports and output ports and thereby increasing efficiency of the network.

The switched data network illustrated in Figure 5 requests permission to transmit a packet through a buffer-less switch **500**. More specifically, the request for permission includes communicating with switch scheduler **510** via signal **REQ 580**. In response, switch scheduler **510** provides one of a grant or a denial of permission via signal **GNT 590**.

The data packet is transferred through the buffer-less switch in an assigned transmission slot. Because there are no buffers in the switch to resolve conflicts, forwarding delays through the switch are fixed. That is, it takes a fixed amount of time for a packet to cross the switch. Being buffer-less does not imply that there can be no storage elements in the switch, it simply means that any switch storage elements that are present do not provide buffering resulting in variable transmission delays through the switch. Thus, any time a portion of a packet is stored in the switch, it is stored for a fixed amount of time before it is forwarded on. That simplifies scheduling of the switch.

An assigned transmission slot is received from the switch scheduler **510** via **GNT 590**. The requests via **REQ 580** and grants via **GNT 590** may be transmitted through separate physical media (one embodiment of which is shown in Fig. 4). A number of different signaling approaches for REQ and GNT signal may be utilized. For example, such signals may be provided on discrete signal wires or be transmitted via the switch itself. In addition, the media used for the requests and grants does not have to match the media of the balance of the network. One of ordinary skill appreciates that any viable communication media may be adapted for the purpose

00540730-033400

described. For example, the media including wire, wireless, optical fiber, or twisted pair are appropriate media for the grant and request lines, or for the network itself.

The nodes of switched data network 500 queue the data packets outside the buffer-less switch 500. For example, node 520, which is optionally a computer system, queues the information to be transferred on the network within a main memory and also within a network interface coupled to the memory. In one embodiment, the memory is a main memory coupled to the network interface and the buffer-less switch 500 via an interconnect such as a bus.

Switch scheduler 510 controls transmit buffer 560, which may be implemented as a queue, and which is coupled to the buffer-less switch 500. The switch scheduler 510 grants requests for transmittal of at least one of a plurality of data packets. In an embodiment, the switch scheduler 510 globally schedules each node coupled to buffer-less switch 500. Thus, for example, if node 520 requests to transmit a packet, the switch scheduler 510 grants the request by assigning a transmission slot to the requesting node 520. All nodes coupled to the buffer-less switch request transmission slots for transmitting through the buffer-less switch 500.

Referring to Figure 6, node 520 is shown in further detail. Node 520 stores a minimal portion of queues 600 within network interface 610, which is within node 520 and coupled to the buffer-less switch 500. Node 520 stores another major portion of the queue within memory 620. In an embodiment, the network interface 610 stores end portions 614 of one or more receive queues 618 and stores leading portions 616 of one or more send queues 622. The network interface 610 holding the leading and the end portions couples to the send queues 622 and the receive queues 618, respectively, via an interconnect 630, the send queues 622 and the receive queues 618 being in memory 620.

The interconnect 630 coupling the network interface 610 and the memory 620 may have unpredictable availability for transfers to and from network interface 610 due to conflicting demands for the interconnect and the scheduling strategy chosen for interconnect 630. That is particularly true if interconnect 630 is a major system input/output bus for the node 520. Thus, placing a minimal portion of the queues 600 in the network interface 610 lessens the probability that delays caused by

unavailability of interconnect 630 will result in delays on network switch 500. Interconnect 630 may also be a point to point connection with predictable availability. If so, delays and unpredictability on interconnect 630 may not be a factor.

5 Preferably, node 520 is one node in a switched data network that includes several network nodes coupled to the network switch. Each node is optionally a computer system including a processor and a memory coupled to the processor or other appropriate system, such as a storage or input/output node. The connection between the nodes and the network switch is optionally a wire, a wireless transmission medium or other appropriate connection depending on system requirements.

Optionally, the buffer-less switch is one of several switches cascaded, forming a multi-stage switch configuration to increase the number of network nodes. A simple embodiment of a multi-stage switch configuration is illustrated in Fig. 16.

15 Referring to Fig. 7, another embodiment of an exemplary network node 700 is illustrated. In one embodiment, network interface card (NIC) 701 of node 700 is based on Active Messages 2.0 and the Virtual Network abstraction (*see generally*, A. Mainwaring: *Active Message Application Programming Interface and Communication Subsystem Organization*. University of California at Berkeley, Computer Science Department, Technical Report UCB CSD-96-918, October 1996; 20 A. Mainwaring and D. Culler: *Design Challenges of Virtual Networks: Fast, General-Purpose Communication*. ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP), Atlanta, Georgia, May 4-6, 1999; B. Chun, A. Mainwaring, and D. Culler: *Virtual Network Transport Protocols for Myrinet*. IEEE Micro, vol. 18, no. 1, January/February 1998, pp. 53-63). This abstraction virtualizes 25 the access points of the network in the form of endpoints. A collection of endpoints forms a virtual network with a unique protection domain. Messages are exchanged between endpoints, and traffic in one virtual network is not visible to other virtual networks. Endpoints are mapped into the address space of a process and can be directly accessed by the corresponding user-level program or kernel program. Thus, 30 user-level communication does not involve the operating system.

NIC 701 holds a small number of active endpoints EP 702. The less active endpoints are stored in main memory 703. The endpoint information stored in the NIC 701 includes pointers to queues in main memory. There are separate queues for the quick channel and the bulk channel. To prevent fetch deadlock of the transfer-acknowledgment protocol, queues come in pairs, that is, there are separate queues for transfers and acknowledgments. There is one pair of queues each for sending and receiving messages over the quick channel. For the bulk channel, there is one pair of send queues, e.g., 705, for each receiving node and one pair of receive queues, e.g., 707 for all sending nodes. Thus, as shown in Figure 7, there are 16 pairs of send queues and 1 pair of receive queues for a 16 port switch. In addition, there is an error queue 709 for reporting erroneous transmissions.

Two types of messages are supported by the illustrated node 700. Quick messages containing a 64-byte payload and bulk messages containing a 1-kByte payload. Figure 4 shows the queues holding the corresponding message descriptors. The bulk and quick packet descriptor formats of the message descriptors are shown in Table 1. While the quick message descriptor contains an immediate payload, the bulk message descriptor contains an immediate payload and an additional payload specified by memory addresses pointing to the source and destination of the transfer. The staging buffers 711 hold that additional payload on its way from and to the main memory. Note that a bulk message descriptor can describe a transfer that includes many bulk packet transfers.

Table 1

Bulk message descriptor		Quick message descriptor	
message type	4 bits	message type	4 bits
source node id	4 bits	source node id	4 bits
source endpoint id	2 bits	source endpoint id	2 bits
source endpoint key	32 bits	source endpoint key	32 bits
destination node id	4 bits	destination node id	4 bits
destination endpoint id	2 bits	destination endpoint id	2 bits
destination endpoint key	32 bits	destination endpoint key	32 bits
immediate payload	44 bytes	immediate payload	64 bytes
source address	64 bits		
destination address	64 bits		
transfer length	32 bits		
	Total: 74 bytes		Total: 74 bytes

Since endpoints are accessed directly by user-level programs, memory addresses specified by the bulk message descriptor are virtual addresses. This requires address translations when message payloads are read from memory by the initiator and written to memory by the target. For this purpose, NIC 701 contains a local translation lookaside buffer (TLB) 713. TLB hits are handled in hardware, while TLB misses are handled in software. Since resolving a TLB miss may take a considerable amount of time, the receiving nodes drops messages that cause TLB misses since such messages could easily flood staging memory.

While several embodiments of various nodes and network interface cards have been described herein, one of skill in the art understands that those embodiment are exemplary only and a wide variety of node designs and network interfaces can be used to practice the various embodiments of the invention described herein.

Referring now to Figure 8, a simple block diagram illustrates an embodiment of a low-latency switch that can be utilized in the embodiments shown in Figures 1 and 4. A low latency communication channel provides the ability to keep latency low for those kinds of communication for which low latency is particularly desirable. One type of communication for which low latency is valuable, besides those mentioned previously in this specification, is remote procedure calls. Communication latency includes sender overhead, transmission time, transport latency and receiver overhead. The low-latency network described herein can reduce communication latency, and, in particular, transmission time.

Low latency is achieved, in part, by allowing a network to lose packets. That way, an optimistic approach can be taken when planning the use of shared network resources such as output ports of a switch. Rather than coordinating and scheduling accesses to shared resources, such as registers, buffers, and, in particular, transmission paths, resources are assumed to be always available. In the event of a conflict, one packet wins and the other ones fail. If transmission fails, it is the sender's responsibility to resend the packet. The lossy network scheme works well in that it saves latency by avoiding time-consuming scheduling operations as long as the network resources are only lightly loaded and conflicts occur infrequently. Thus, it is preferable that a lossy network is designed in a way that the switches and links are not

highly loaded, by providing, e.g., excess bandwidth. Excess bandwidth helps keep dropped packets to a minimum.

A lossy network is particularly attractive since it allows one to build simple and fast switches such as the switch illustrated in Figure 8. Although a 2X2 switch is illustrated for ease of understanding, the concepts described herein associated with a lossy switch can be incorporated into any size switch. No time-consuming arbitration or scheduling of its data paths is required. Packets are forwarded on a first come first served basis. Thus, as shown in Figure 9A, packet B is dropped because it arrived at the output port selector circuit later than packet A. If packets do happen to collide, one packet wins and the other packet(s) are dropped. Thus, as shown in Fig. 9B, packet A is chosen as the loser based on some simple algorithm such as a random or round robin selection. More sophisticated algorithms can be chosen such as selecting the winner according to a fairness criteria having an objective to allocate the same amount of output port bandwidth to each input port on the switch. Any approach used to choose a winner should preferably not add any more than necessary to latency.

Lossy communication also makes it possible to use simple and fast buffering schemes in the sending and receiving nodes of the network. Referring to Figure 10, assume that the sender and the receiver are either a user program, a systems program, or a transmission protocol. Figure 10 again illustrates a buffer-free 2X2 switch. Assume that node 0 is sending a packet. To send a packet, a node 0 writes a packet into send register 1012. Node 0 then polls a status register 1014 until it becomes valid. Once the status register is valid, it will indicate whether the transmission was successful. If the status register indicates that the transmission was unsuccessful, the sender has to resend the packet by writing the packet into send register 1012. Because low latency communication is typically synchronous in that a sender cannot proceed until it is known that the transmission was successful, the sender can be put in charge of doing the retransmission if necessary. Successful and unsuccessful transmission can be determined with the help of an acknowledge packet (ack) or no acknowledge packet (nack), respectively, or a timeout mechanism in which the sending node waits a predetermined amount of time to see if an acknowledge indicating a successful transmission is received. If not, the sender assumes an error. When the target is node 1, the status register 1014 may receive an ack written into the node 1 send register



5

15

25

5 A timeout mechanism can inform the sender accordingly. A more sophisticated approach reports errors to the sender to allow the system to better determine the cause of packet loss. In any case, the network does not have to be able to retransmit erroneously transmitted packets, as that task is left to the sender. In fact, the task may be left to kernel software or application or user programs that made the transfer.

No intermediate buffers are needed between the sender and the receiver, as are typically found in other switching networks. If conflicts occur, rather than buffering a packet, packets are simply dropped. As a consequence, no buffering or buffer management including flow control is needed.

While the quick channel has minimum scheduling, one implementation for the  
30 bulk channel relies on pipelining to increase throughput. Pipelining is a technique to  
increase throughput by overlapping the execution of multiple operations. A pipeline

breaks the execution of an operation into several steps also called pipeline stages. Overlapped execution is achieved in that each stage operates on a different operation. In its simplest form, a pipeline has a fixed number of stages of equal length. One advantage of applying pipeline techniques to computer networks is that they simplify design of the computer network. Referring to Fig. 11A, three sequential operations are shown OP1, OP2 and OP3. When pipeline techniques are used, portions of those operations can be overlapped as shown in Fig. 11B. Each operation shown is divided in three stages S0, S1 and S2. As can be seen, stage S1 from OP1 can be overlapped with stage S0 from OP2. The overlapping of the other stages is readily apparent from Fig. 11B. Because the operations no longer have to be executed serially, but can be executed at least partially in parallel, the execution rate is improved. The pipelined execution shown in Fig. 11B results in three times the throughput of the serial execution shown in Fig. 11A.

While the pipeline techniques are applicable to the bulk channel disclosed herein, the pipeline techniques described herein for a network are applicable to any network that can advantageously exploit the teachings herein regarding pipelined networks. Consider for example, a switched network with fixed forwarding delays that executes remote DMA write operations. The node that sources the data is called the initiator and the node that sinks the data is called the target.

In one embodiment, a pipeline implementation of a network includes the following four stages. An arbitration stage (ARB) is the stage in which initiators request routing paths and an arbiter calculates a schedule based on the routing paths requested by the initiators. A transfer stage (TRF) follows an arbitration stage. During the transfer stage, a transfer packet containing the data is sent from the initiator to the target. An acknowledge stage (ACK) follows the transfer stage. During the acknowledge stage the target returns an acknowledge packet containing a delivery report to the initiator. Finally, in this embodiment a check stage (CHK) follows the acknowledge stage in which the acknowledge packet is checked by the initiator to determine whether the operation succeeded. More stages might be required, for example, to transmit the transfer and acknowledge packets described.

In one embodiment packet size is fixed. If the remote DMA operation wants to transfer more data than fits into a single transfer packet, multiple transfer packets and with it multiple operations are needed. Fixed packet size greatly simplifies scheduling of the network. A pipelined network executes operations in bounded time.  
 5 That simplifies the design in at least two areas, error detection and switch scheduling.

A pipelined network simplifies detection of lost packets. Networks typically have some degree of unreliability, in that a packet can be lost or erroneously transmitted. To detect this, handshaking protocols are used. Basically, such protocols confirm the receipt of a transfer packet by sending an acknowledgment packet back to  
 10 the initiator. If the transmission paths of the network as well as the network interfaces are pipelined, the initiator can wait for a fixed amount of time, check for the arrival of an acknowledge packet and determine whether transmission succeeded.

In comparison with present network implementations that exhibit variable and unbounded transmission delays, the pipelined network provides several advantages.  
 15 The number of outstanding operations and, with it, unacknowledged packets is fixed. With variable and unbound transmission delays, that number varies and can be quite large. Since state has to be stored for each unacknowledged packet, a large state memory and possibly sophisticated state management is required. Additionally, messages on the pipelined network are delivered in order. To guarantee progress, a  
 20 network with variable delays often delivers packets out of order. That complicates bookkeeping of unacknowledged packets and assembling packets into larger data entities.

In a preferred embodiment, the pipelined network described herein has fixed forwarding delays for all transmission paths. It is, therefore, particularly well suited  
 25 for small networks with a limited diameter and with a small number of nodes with a single switch connecting the nodes. It is also possible to cascade switches to increase the number of nodes that can be connected.

Referring now to Figure 12, a packet flow diagram illustrates an embodiment of a synchronous pipelined network in which boundaries of all stages are aligned.  
 30 Figure 12 demonstrates a plurality of stages, including an arbitration stage 1210, a transfer stage 1212, and an acknowledge stage 1214. As shown, each of the stages

1220 has a fixed time relation to each other stage. The stages are shown to have equal length, however, one of skill in the art appreciates that the length of the stages optionally is variable depending on design requirements. Also, the number of stages may vary depending on design requirements. For example, the transfer stage could be split up into several stages. Figure 12 shows a check stage 1216, as an optional stage. The check stage 1216 provides an optional stage in which sending nodes check if transmission of a sent packet was successful. The check stage is optional in that it can be omitted if the acknowledge stage already checks for successful transmission. Figure 12 illustrates transactions occurring between two nodes of a network. Other transactions between other nodes in a switched network system may also be occurring at the same time.

As can be seen in Figure 12, the transfer stage 1212 during which a packet is being transferred across the network can be overlapped with a subsequent arbitration stage 1210. In fact, all four stages can be overlapped. One approach to providing overlapping operations can utilize a network such as the one shown in Figure 4. Referring again to Fig. 4, assume that that bulk channel 450 is part of the pipelined switched network. The arbitration stage can utilize the quick channel 460 to send request packets and grant packets. During an arbitration stage, a vector of requests can be sent from a requesting node to the arbiter shown as bulk switch scheduler 440 and bulk switch scheduler 440 can send a grant to the requesting node. To avoid potential conflicts between arbitration stage packets (request and grant packets) and other traffic on quick channel 460, a scheme as described with relation to quick channel 460 can be used such that request packets from the nodes during the arbitration stage can be forwarded directly from the input ports 422 of quick channel switch 460 to the switch scheduler 440 without passing through the switching fabric of quick channel switch 460. Grant packets are given higher priority than other packets when they are forwarded from bulk scheduler 440 to output ports 432 to avoid conflicts with other quick channel traffic. That avoids collisions within the switching fabric and the potential of dropping of request and grant packets.

In the embodiment shown in Figure 4, the nodes 430 send acknowledge packets during the acknowledge stage to nodes 420 on quick channel 460 in response to data transferred during the transfer stage. Those acknowledge packets are

transferred within the switch fabric of quick channel 460. The timing of sending acknowledgment packets can be chosen such that collisions with request and grant packets are avoided. If nodes 420 simultaneously send acknowledgment packets in response to transfer packets sent during the previous bulk frame, and if the  
5 acknowledgment packets are sent at a different time than the request and grant packets, it is guaranteed that the acknowledge packets can not collide in quick channel switch 460 with either the request and grant packets.

That can be accomplished as follows. Assume the nodes and the switch use a common schedule to schedule the transmission of request, grant and acknowledgment  
10 packets. There are fixed times relative to the bulk frame when those packets are sent. For example, assume a bulk frame takes 1024 time units. Also assume that the request packets are transferred from the initiator nodes to the switch scheduler at time 1, the grant packets are transferred from the switch to the initiator nodes at time 512, and the bulk channel acknowledge packets are transferred from the target nodes to the  
15 initiator nodes at time 256. Since the packets are sent at different times, they cannot collide with each other.

There could be collisions of the three types of packets mentioned with regular packets sent over the quick channel. As previously described, the request packets will not collide with regular packets since they are taken off the network at the input ports  
20 of the switch, from where they are forwarded to the arbiter, and, therefore, do not pass through the switching fabric where collisions could occur. The grant packets are forwarded from the arbiter to the output ports of the switch where they are injected into the network. Logically, there is a separate input port connected to the arbiter. Grant packets can collide with regular packets. If that happens, grant packets win and  
25 regular packets lose, as previously stated. Since the nodes know the time when the grant packets are sent, they could avoid conflicts by not sending regular packets in the corresponding slot.

Acknowledge packets are handled similarly to the grant packets. If there is a collision with a regular packet, the grant packet wins and the regular packet loses.  
30 Note that in some implementations, there should not be any regular packet present in the network when acknowledge packets are transmitted. Assuming every node sends

an acknowledge packet, and acknowledge packets are sent at the same time, there can only be regular packets in the network in case of an error or a misbehaving node.

The acknowledge packets can be forwarded through the quick switch in a conflict-free manner. The settings of the quick channel switch used for forwarding the acknowledge packets correspond to the inverted settings of the bulk channel switch used for forwarding the corresponding transfer packets – it is the direction of transfers that has been reversed. E.g., if the transfer packet was transferred from input port 1 to output port 2 of the bulk channel, the acknowledge packet needs to be forwarded from input port 2 to output port 1 of the quick channel switch.

In addition, because the pipeline is synchronized, the quick switch can send a special packet once per bulk frame to each node for synchronization purposes. The grant packet, which may be sent at a fixed time in the frame (e.g. at time 512), can be used for synchronization purposes by the nodes. The quick channel switch transfers a grant packet to every node once per bulk frame. All nodes implicitly know the time relative to the bulk frame that the grant packet is sent. Therefore, the receipt of a grant packet by a node can be used as a time reference and the node can derive the beginning of a frame from this reference. The grant packet may also be used to supply the node with a unique identifier. In that case, each grant packet contains a unique identifier which corresponds to the number of the output port through which the grant packet was sent. During node initialization, the node listens to grant packets and uses the supplied identifier as its node identifier which is used by all communication for identifying a node.

The pipelined network may include a flow control mechanism. In one embodiment, an arbiter, on receiving a request for a particular output port, queries the node at the output port for its availability to receive a packet or packets. The node replies with a go/no-go to the arbiter as to its readiness as a form of simple flow control. The arbiter then allocates the output port according to availability and other criteria it uses in its arbitration scheme. The packets that include flow control information are also preferably transferred over the quick channel.

In typical networks, each node is typically both an initiator node and a target node. That is, each node is generally coupled to both an input port and an output port.

That allows, in one embodiment, for the flow control information to be included in the request packet in the form of a bit vector that specifies which initiator may send packets to the node (as a target) that is sending the request packet. That flow control information may be based on the state of queues, which a node may have dedicated to a particular initiator. Thus, if the queue holding data from initiator 1 is full, the bit vector would indicate that the node was unable to accept any further data from initiator 1.

Note that the bit vector for flow control purposes may also be used to ignore a node that is, e.g., determined to be broken. In that way, a misbehaving node can be ignored. Similarly, the ports on the quick channel switch may be enabled and disabled using an enable vector sent from the various nodes as part of the grant packet.

Note that the length of the transfer stage may provide more time than is necessary to complete the arbitration stage and the acknowledge stage. The check stage is performed in the node and therefore generally does not interfere with other pipeline operations.

Because it can be advantageous in terms of bulk channel speed to implement the bulk channel with a flow through switch that does not need to extract data from the packets, arbitrating over the quick channel is an advantage. The quick channel, on the other hand, does extract data from the data packets to select destination ports, for example, and thus can be advantageously used for arbitration as well. If the bulk channel carried arbitration traffic as well, one could intersperse request and grant packets between packets sent during the transfer stage. But that would separate the transmission of the request packet and the grant packet by one bulk frame, possibly requiring one more pipeline stage before the corresponding data could be sent in the transfer stage. Note that in some embodiments, the arbiter also has to determine at least minimum flow control information from the targets. For the same reasons, transmission of the acknowledge packet in response to a packet sent over the bulk channel during the transfer stage is preferably done over the quick channel.

Depending on the type of scheduling that is used in a pipelined network implementation, conflicts arise if multiple packets are to be transferred over a



5

10

20

25

Thus, the collisions are detected by the handshaking and the system resends data in response. Applied to the example of a switched network, the initiator detects the loss of a packet if it does not receive an acknowledge packet a certain amount of time after the transfer packet was inserted into the pipeline. That scheme to detect collisions can  
5 be attractive if collisions are infrequent and if end-to-end latency, as well as the time taken to calculate a schedule, is to be kept as short as possible.

Thus a pipelined network has been described that may be implemented as a switched, ring-based or a bus-based network or a combination. The network  
10 pipelining techniques are applicable to any network that can overlap pipeline stages to achieve greater throughput on the network.

The embodiments of the data networks, computer systems, methods and  
switches described above are presented as examples and are subject to other variations in structure and implementation within the capabilities of one reasonably skilled in the art. The details provided above should be interpreted as illustrative and not as  
15 limiting. For example while the various embodiments have generally shown single switch stages, any of the switches shown herein can be cascaded into multiple switch stages and/or be cascaded with other switched or bused networks. Other variations and modifications of the embodiments disclosed herein, may be made based on the description set forth herein, without departing from the scope and spirit of the  
20 invention as set forth in the following claims.